Aviation Human Factors Division
Institute of Aviation

University of Illinois
at Urbana-Champaign
1 Airport Road
Savoy, Illinois 61874

On The Independence of Compliance and Reliance: Are Automation False Alarms Worse Than Misses?

Stephen R. Dixon,
Christopher D. Wickens,
& Jason S. McCarley

Technical Report
AHFD-05-16/MAAD-05-04

March 2006

Prepared for

Micro Analysis and Design
Boulder CO

Contract ARMY MAD 6021.000-01

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
MAR 2006		**00-00-2006 to 00-00-2006**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
On The Independence of Compliance and Reliance: Are Automation False Alarms Worse Than Misses?	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Illinois at Urbana-Champaign,Aviation Human Factors Division,Savoy,IL,61874**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

(a) Objective: Participants performed a tracking task and system monitoring task while aided by diagnostic automation. The goal of the study was to examine operator compliance and reliance as affected by automation failures, and to clarify claims regarding independence of these two constructs. (b) Background: Background data revealed a trend towards non-independence of the compliance-reliance constructs. (c) Method: Thirty-two undergraduate students performed the simulation that presented the visual display and collected dependent measures. (d) Results: False alarm prone automation hurt overall performance more than miss-prone automation. False alarm prone automation also clearly affected both operator compliance and reliance, while miss-prone automation only appeared to affect operator reliance. (e) Conclusion: Compliance and reliance do not appear to be entirely independent of each other. (f) Application: False alarms appear to be more damaging to overall performance than misses, and designers must take the compliance-reliance constructs into affect.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**14**	

Abstract

(a) **Objective**: Participants performed a tracking task and system monitoring task while aided by diagnostic automation. The goal of the study was to examine operator compliance and reliance as affected by automation failures, and to clarify claims regarding independence of these two constructs. (b) **Background**: Background data revealed a trend towards non-independence of the compliance-reliance constructs. (c) **Method**: Thirty-two undergraduate students performed the simulation that presented the visual display and collected dependent measures. (d) **Results**: False alarm prone automation hurt overall performance more than miss-prone automation. False alarm prone automation also clearly affected both operator compliance and reliance, while miss-prone automation only appeared to affect operator reliance. (e) **Conclusion**: Compliance and reliance do not appear to be entirely independent of each other. (f) **Application**: False alarms appear to be more damaging to overall performance than misses, and designers must take the compliance-reliance constructs into affect.

Introduction

For the past few decades, designers have attempted to reduce operator workload levels by introducing automated aids that assist or replace human functions. One example is the diagnostic aid that alerts operators to potential problems in the environment. The current study examines the effects of diagnostic aids on benefit human performance in a supervisory control situation.

Imperfect Automation

It is sometimes tempting to assume that automation is a panacea for reducing workload levels and improving performance, but not all forms of automation are perfectly reliable. This is particularly true of diagnostic automation, which must make diagnoses and long-range predictions in a world of imperfect probabilistic information (Wickens & Dixon, in press). Diagnostic automation designed to detect a state of the world can produce two forms of errors—false alarms and misses. A diagnostic aid's performance can be measured using signal detection theory (Green & Swets, 1988). The automation can have a liberal criterion (or beta) or a conservative criterion; that is, it can either commit more false alarms or more misses, respectively, or it can be neutral.

Since beta is typically set at the discretion of the system designer, it is critical for designers to understand the effects of different forms of automation errors before implementing a new automated system. This includes understanding the effects of the drivers of optimal beta; that is, how the probability of a signal, the costs of error, and the payoffs of correct responses interact with the operators' determination as where to set their criterion. The goal of the current study was to provide data that allow system designers to better make these judgments during the design process.

Compliance and Reliance

Recent theory has postulated that automation false alarms and misses affects have qualitatively different effects on operator dependence (Meyer, 2001; 2004), respectively. Compliance is what the operator does when the automation diagnoses a signal in the world, while reliance is what the operator does when the automation diagnoses noise in the world. An increase in false alarms is posited to reduce compliance, resulting in longer response times to automation alerts, or, in extreme cases, a tendency to disregard of those alerts entirely – the "cry wolf" effect. (Dixon & Wickens, in press; Wickens, Dixon, Goh, & Hammer, 2005). An increase in the automation's miss rate reduces reliance, causing the operator to allocate more attention to monitoring the raw data behind the automation in order to catch the possible automation misses. This diverts attentional resources from the concurrent task, and a subsequent deterioration in performance in that task (Dixon & Wickens, in press; Wickens, Dixon, Goh, & Hammer, 2005).

Previously, it had been implied, although not explicityly, that compliance and reliance may be independent constructs (Meyer, 2004); that is, an increase in false alarms should only affect compliance and an increase in misses should only affect reliance, and indeed this could be seen to be an optimal cognitive response.

Dixon and Wickens (in press) reported data partially consistent with this assertion. They had pilots fly a simulated unmanned aerial vehicle (UAV) mission, which consisted of tracking

the UAV through a series of waypoints while searching for targets of opportunity along the way. Concurrently, the pilots were responsible for monitoring a set of four system gauges for possible system failures, aided by the implementation of a diagnostic aid that sounded an auditory alert when it determined (correctly or incorrectly) that a system failure had occurred. Importantly, the system gauges were perceptual in nature (green and red zones) and required very little, if any, cognitive resources to analyze the raw data behind the alarms. The data revealed that increasing the automation miss rate only affected measures of reliance, while increasing the automation false alarm rate appeared to affect both compliance and reliance. However, due to the low power and sensitivity of the concurrent task measures, the authors were unable to find strong statistical evidence of the non-independence of the constructs.

Wickens, Dixon, Goh and Hammer (2005) replicated a portion of the previous experiment, and added eye tracking data to their analyses. They found behavioral data consistent with the previous findings, and their measures of visual scanning provided further evidence for the possible non-independence of reliance and compliance. Specifically, the investigators noted that high false alarm rates induced a significant shift of attention toward the raw data in the alerted domain, and therefore away from concurrent tasks. This should have been solely a symptom of a high automation miss rate according to the independence model. However, the same issues of low statistical power prevented the authors from making strong claims of non-independence.

Wickens, Dixon and Johnson (2005) repeated a similar version of the UAV paradigm, but instead provided an unreliable diagnostic aid to the more difficult target search task, while also providing a perfectly reliable aid to the system gauge task. They found that the disruptive effects of automation false alarms on the concurrent task was at least as strong as the automation misses, yet another finding inconsistent with the independence model of reliance-compliance. Similar effects have been found recently in a multiple-UAV paradigm (Levinthal & Wickens, 2005).

Collectively, these studies suggest that false-alarm prone automation may be, overall, more disruptive of multi-task performance than miss-prone automation (Bliss, 2003), because of the former's effect on concurrent task performance as well as automated task performance. However, this conclusion is based on only marginally significant performance trends with a low power measure (Dixon & Wickens, in press), or on visual scanning data (Wickens, Dixon, Goh & Hammer, 2005). Furthermore, some of these findings may be a function of the perceived cost of false alarms and misses on total human-machine output.

The Current Study

The current study addresses the weaknesses of the experiments described above by providing a continuous and sensitive measure of concurrent task performance, hence allowing greater statistical power and experimental control. Furthermore, in the automated task, participants encounter over 35 examples of one class of automation error (misses or false alarms), whereas in the prior studies the disparity was far less (typically 2 or 3 of one class). Given these changes, the current study provides a stronger opportunity to examine the relationship between reliance and compliance in a way that was not available in the previous studies. To the extent that reliance and compliance *are* independent constructs influenced by automation misses and false alarms respectively, then performance in the concurrent tracking

task during non-alert periods should be equivalent between a perfectly reliable automation condition and a FA-prone condition, since the operator should not be monitoring the systems gauge if there is no automated alert (i.e. unaffected reliance).

The second important addition to this study is that it continues to extend the conclusions beyond the relatively simple perceptual monitoring task used in the early UAV studies to one with highly demanding cognitive elements—a manipulation also subsequently done by Levinthal and Wickens (2005) and Wickens, Dixon, and Johnson (2005).

The current study involved two concurrent tasks: a continuous compensatory tracking task and a cognitively demanding systems monitoring task. In the latter, participants were required to calculate current values and report when the needle of the system gauge exceeded a certain acceptable range. Some participants performed the task unaided, while others performed the task with the aid of an automated diagnostic system that was either perfectly reliable, FA-prone, or miss-prone.

We hypothesized that: a) perfect automation would benefit both the tracking task and systems monitoring task; b) the system prone to automation misses would harm the tracking task due to a reduction in operator reliance, causing a shift of attention away from the tracking task in order to catch the automation misses in the systems monitoring task; c) an increase in the automation miss rate should not affect measures of operator compliance (e.g., speed of response to an alert); d) the system prone to automation false alarms would harm the system monitoring task due to a reduction in operator compliance; e) automation false alarms would also harm the tracking task, even when the alarm is silent, due to a reduction in operator reliance; and f) as a consequence of the previous hypotheses, automation false alarms would be more harmful to overall performance than automation misses.

Method

Participants

Thirty-two undergraduate students from the University of Illinois participated in the experiment, and were paid $9 per hour, plus performance bonuses. Participants were made aware of the incentives, and were told that the two tasks should receive equal priority.

Apparatus and Stimuli

The experimental simulation ran on a Dell GX270 computer, with a 21” Dell monitor, using 1280x1024 resolution. Figure 1 presents a sample display for the experiment, with verbal explanations for each display window and task.

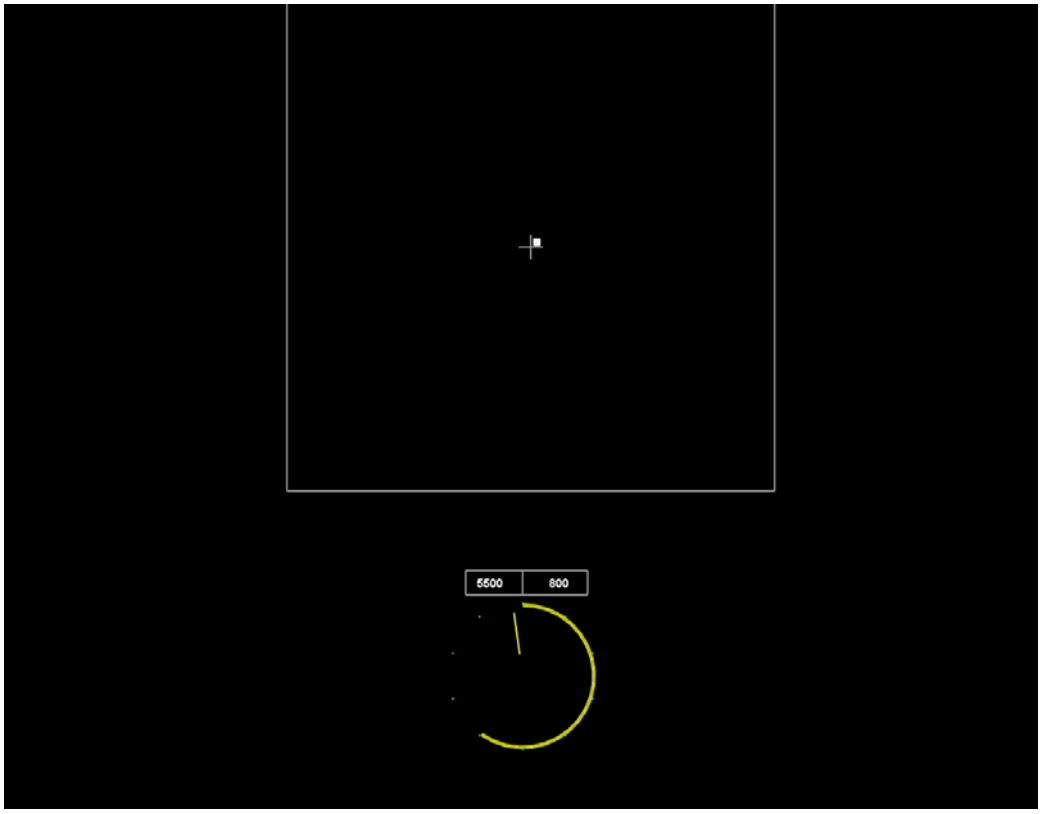

Figure 1. Sample screenshot of experimental display.

The experimental display was subdivided into two areas of interest, separated by approximately 12 degrees of visual angle (center to center). In the top portion of the display was the two dimensional tracking task. Participants used a joystick with first order dynamics to track the target disturbed by a quasi random input with a bandwidth of 0.43 Hz. A negative drift in error was added so that if the participant did not exert feedback on the joystick, the ball would quickly float towards the outer edges of the box.

The system monitoring gauge was located in the bottom display, and represented the value of a generic real-world variable (e.g., altitude). The gauge had ten small white ticks spaced equidistantly around the outside of an imaginary circle. A yellow bar that "filled" the outside of the gauge denoted units of 1000, 2000, etc. A yellow needle that rotated around the inside of the gauge denoted units of 100, 200, etc. Thus, the value of the sample gauge depicted in Figure 1 is approximately 5975. The yellow needle was driven by the sum of four sine waves ranging in bandwidth from 0.04 Hz to 0.43 Hz. The yellow bar "filled" the gauge in a linear fashion, according to whatever the current value was, as dictated by the random movements of the needle.

Above the gauge were located two white boxes with white numerical values. The number in the left box denoted the ideal value for a 'safe' system. The number in the right box denoted the range of 'safety' for the system. Thus, for the example shown in Figure 1, the participant had to keep the gauge within 5500 +/- 800. If the gauge went out of this range, it denoted a system failure (SF). If a SF occurred, the participant was expected to press a button on the keyboard as quickly as possible. When a SF occurred, the needle stayed out of the acceptable value range

until it was either detected or the trial ended. The SF task was purposely designed to be a challenging task that required both visual and cognitive (working memory) resources. While in a real-world application, the system gauge might be considered a poorly designed gauge, in that it is difficult to read, it was designed for this experiment as much for its theoretical value as for its practical value, and in recognition of the fact that many real-world gauges indeed *are* poorly designed.

For some participants, performance on the systems monitoring task was aided by automation. The automated aid sounded an auditory alert (i.e. a synthesized human voice pronouncing the word "one") when a SF occurred. The automation, expressed in the framework of signal detection theory (SDT), could provide hits (alarm with true SF), misses (no alarm with true SF), false alarms (alarm with no SF), or correct rejections (no alarm with no SF).

Trials

There were 100 trials that each lasted exactly 30 seconds. At the beginning of each trial, the target value (in the left numeric box above the SF gauge) changed to a new random value between 1000 and 9000, rounded to the nearest 100. The target range (in right numeric box above the SF gauge) changed to a new random value between 100 and 900, rounded to the nearest 100. The SF gauge itself reset to the target value and then immediately began oscillating.

A system failure occurred on 50 trials, with SF and non-SF trials randomly ordered. SFs (and automation false alarms) always occurred within a temporal window beginning 5 seconds and ending 12 seconds from the start of the 30-second trial interval, thus giving the participant at least 18 seconds to detect the failure. There was never more than one SF or alert from the automation per trial. Trials lasted the entire 30 seconds, regardless of whether or not an SF occurred or was detected. During each trial, the participant was allowed to make only one SF response (e.g. if they responded "yes" before a SF actually occurred, then it would be classified as an operator false alarm), and were not allowed to retract a response once executed. Once the 30-second trial ended, participants were no longer able to respond to that particular trial. At the end of each trial, the screen either flashed green or red to inform participants whether their response was correct or incorrect, respectively.

Procedure and Design

After filling out a consent form and reading the instructions, each participant completed 20 practice trials followed by 80 experimental trials. There were four experimental conditions: a) Baseline condition (no automated aid), b) A100 condition (40 hits, 0 FAs, 0 Misses, 40 CRs), c) FA60 condition (40 Hits, 32 FAs, 0 Misses, 8 CRs), and d) M60 condition (8 Hits, 0 FAs, 32 Misses, 40 CRs). Participants were told that the automation would either be perfectly reliable or "not perfectly reliable", and in the latter case, which way the automation criterion would be set. Thus, it can be assumed that the participants were immediately aware of the potential for automation failures, as well as which type of failure they would encounter.

Results

One subject in the M60 condition was dropped due to unusually poor performance levels (beyond the third standard deviation below the group mean) in the tracking task. For the most

part, analysis entailed a one-way omnibus ANOVA, followed by three planned comparisons: a) Baseline vs. A100, b) Baseline vs. the combination of FA60 and M60 in a planned comparison (i.e. weights of -1, 0.5, 0.5), and c) FA60 vs. M60. Because only three a priori comparisons were made, familywise error rates were not adjusted (see Keppel, 1982, for more details). Any post-hoc tests used a Bonferroni correction.

Tracking Error

Tracking error was calculated only during the period of time between the beginning of a trial and the onset of either a system failure or an automation false alarm, since this was the period of time where variations in attentional reliance caused by the different conditions were expected. These data represented between 5-12 seconds of time at the beginning of each trial. Figure 2 presents these data as a function of condition, using the solid black bars.

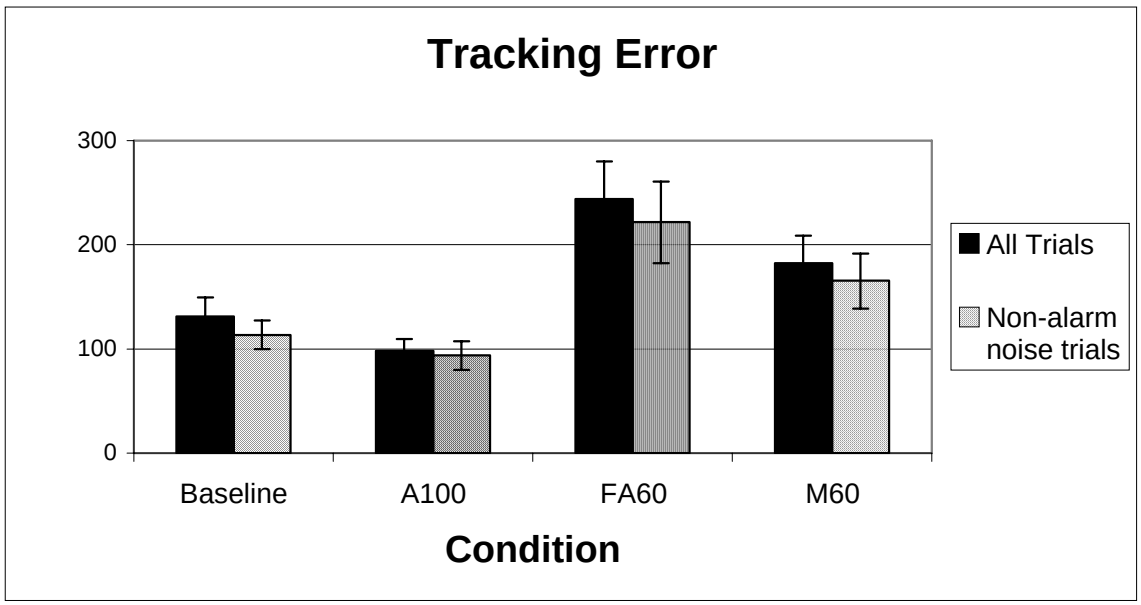

Figure 2. Tracking error as a function of condition. The solid black bars represent all trials in the experiment, while the diagonal pattern bars represent only non-alarm noise trials. SE bars are included.

A one-way ANOVA revealed a reliable main effect of condition, $\underline{F}(3,27) = 6.64$, $p < .01$. Planned comparisons showed that automation ($M = 131$) may have improved performance over Baseline ($M = 98$) at a level approaching significance, $\underline{t}(14) = 1.53$, $p = .07$. The Baseline condition showed better performance relative to the average of the two unreliable conditions, $\underline{t}(14) = 2.55$, $p = .01$, while the difference between the FA60 ($M = 243$) and M60 conditions ($M = 182$) was not statistically significant, $t(13) = 1.32$, $p > .10$.

Effect of False Alarms on Reliance

Because tracking error was measured during the time *before* a SF or alert occurred, we can assume that any performance deficits in the tracking task for the FA60 condition indicate

a reduction in operator reliance. That is, such deficits would demonstrate that the operator was putting attentional resources into the SF task even when there was no alert, causing an increase in tracking error. A separate analysis done only on trials in which there was no SF and the automation was silent revealed the same results. These data are shown with the hatched bars of Figure 2. A one-way ANOVA on these data revealed a main effect of condition, $F(3,27) = 5.09$, $p < .01$, and a post hoc comparison between the FA60 and the A100 conditions, $t(14) = 3.78$, $p < .01$, provides clear evidence that operator reliance was reduced.

SF Detection Rate

There were no significant differences found in operator beta across conditions, $F(3, 27) = 2.3$, $p > .10$. For all other detection rate analyses, the signal detection measure *d'* was used. Perfect scores (e.g. zero misses or false alarms) were adjusted by assuming ½ of a miss or FA.

A one-way ANOVA revealed a main effect of condition, $F(3, 27) = 8.84$, $p < .001$. Planned comparisons revealed no significant difference between the Baseline condition ($M = 3.03$) and A100 condition ($M = 3.20$), $t(14) < 1.0$. The Baseline condition produced performance better than the average of the two unreliable conditions, $t(14) = 2.43$, $p = .01$. The FA60 condition ($M = 2.04$) performed worse than the M60 condition ($M = 2.61$), $t(13) = 3.08$, $p < .01$. Post hoc tests revealed that the Baseline condition performed better relative to the FA60 condition, $t(14) = 3.15$, $p < .01$, but did not differ significantly from the M60 condition, $t(13) = 1.38$, $p > .10$.

Further analysis was done in the automation conditions to determine the likelihood of a yes/no operator response based on the type of automation response, helping to shed light on operator agreement or disagreement with the automation. Table 1 presents these data.

Table 1. Operator response as a function of automation accuracy. Operator agreement rates are shown. The first and third columns tend to be measures of compliance, while the second and fourth columns tend to be measures of reliance.

	Automation Correct		**Automation Incorrect**	
	Auto Hit	**Auto CR**	**Auto FA**	**Auto Miss**
	Compliance	**Reliance**	**Compliance**	**Reliance**
A100	0.96	0.93	*	*
FA60	0.93	0.92	0.35	*
M60	1.00	0.82	*	0.05

These data indicate that operators generally agree less when automation is wrong than when it is right. When automation is right, compliance is appropriately lowered in the FA60

condition (*M* = .93) compared to the M60 condition (*M* = 1.00), t(14) = 3.75, p < .01, and reliance is appropriately lowered in the M60 condition (*M* = .82) compared to the FA 60 condition (*M* = .92), t(13) = 2.14, p < .05. When automation is wrong, the level of agreement is much less (override the faulty automation) when a miss-prone system makes an error (it misses a true signal: reliance *M* = 0.05) than when a false-alarm prone system makes an error (it alerts a noise trial: compliance *M* = 0.35). Thus participants monitor the raw data in a miss-prone system, more than they check the faultiness of an alert, in the FA-prone system.

SF Response Times

SF response times are presented in Figure 3. The solid black bars represent all trials in the experiment while the hatched bars represent *only true-alarm signal* trials (to be discussed below). Note that the Baseline condition did not have any true alarms, but the corresponding trials were included as a measure of what "baseline" results were.

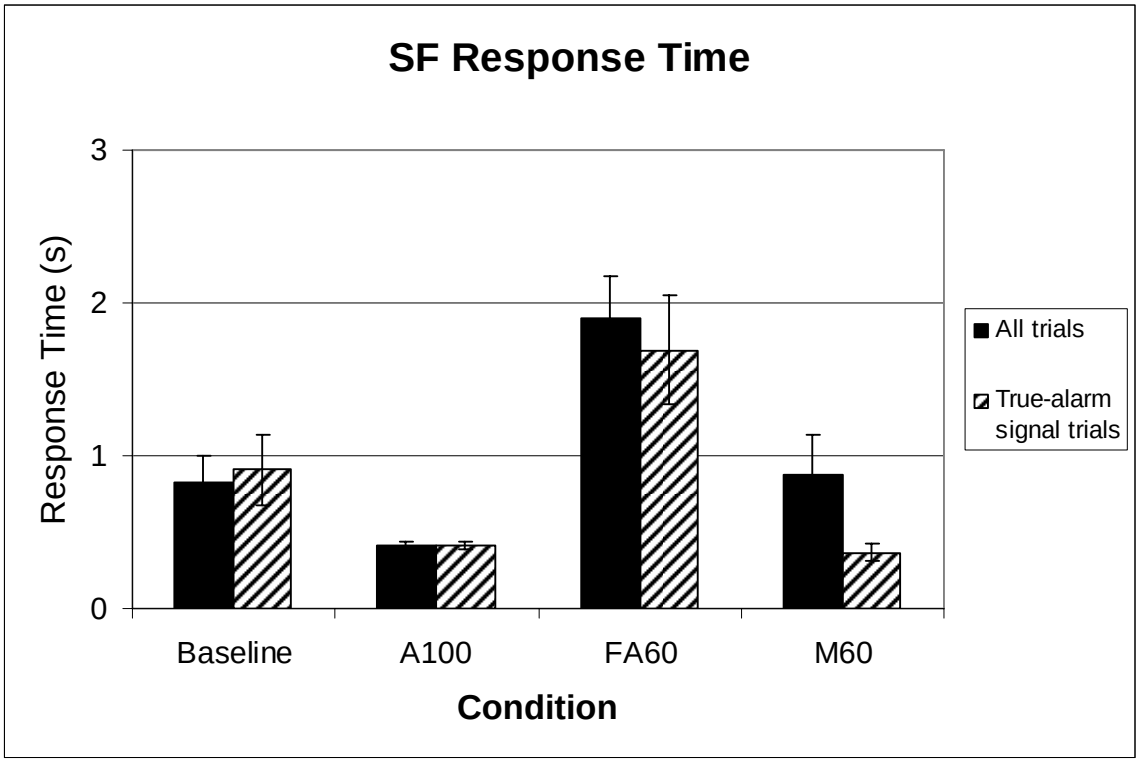

Figure 3. SF response times as a function of Condition. The solid black bars represent all trials in the experiment, while the diagonal pattern bars represent only true-alarm signal trials. SE bars are included.

A one-way ANOVA on the data for all trials revealed a main effect of condition, F(3, 27) = 9.85, p < .001. Planned comparisons revealed that participants in the Baseline condition (*M* = 0.82 s) performed more poorly than those in the A100 condition (*M* = 0.42 s), t(14) = 2.26, p < .05, but better than the average of the two unreliable automation conditions, t(14) = 2.48, p = .01. The FA60 condition (*M* = 1.91) performed much more poorly than the M60 condition (*M* = 0.88), t(13) = 2.71, p < .01. Post hoc tests revealed that performance in the Baseline condition

was better than in the FA60 condition, $t(14) = 3.35$, $p < .001$, but was not significantly different from that in the M60 condition, $t(13) < 1.0$.

Effect of Misses on Compliance

The following analysis was done to determine if automation misses had any effect on operator compliance. Applying the logic used to examine accuracy, if compliance in the M60 condition was perfect, then on trials where the automation alert sounded, the response times to the SFs should have been equivalent to those in the A100 condition; that is, the operators would have known that the automation did not commit false alarms and that when it sounded, it was always correct. To test the effects of compliance, data for trials on which an alarm occurred were analyzed separately.

A one-way ANOVA on these SF response times revealed a main effect of condition, $F(3, 27) = 7.54$, $p < .01$. Planned comparisons between the M60 condition ($M = 0.37$) and A100 condition ($M = 0.41$), $t(13) < 1.0$, revealed that the misses in the M60 condition did not appear to affect operator compliance at all, and that performance in the M60 condition following a true alarm was better than performance in the Baseline condition ($M = 0.91$), $t(13) = 2.10$, $p < .05$. As expected, the FA60 condition ($M = 1.69$) did degrade operator compliance, $t(14) = 3.53$, $p < .01$.

Discussion

Previous studies have indicated that automation false alarms and automation misses have qualitatively different effects on operator performance (Meyer, 2004; Dixon & Wickens, in press); that is, automation false alarms tend to adversely affect operator compliance, while automation misses tend to adversely affect operator reliance. The current study was able to provide a stronger opportunity to examine the degree of independence of the reliance-compliance constructs implied by Meyer (2001, 2004) and expanded upon by Dixon and Wickens (in press).

The current study replicated the finding that perfect automation is beneficial to overall human-automation performance, as predicted in Hypothesis A. Hypothesis B predicted that the miss-prone automation would harm the tracking task by causing operators to shift attention away from the tracking task in order to catch the potential automation misses. This proved to be correct, as the performance in the miss-prone condition suffered relative to the perfectly reliable condition, matching previous findings by Dixon and Wickens (in press). Consistent with Hypothesis C, the data showed that automation misses had no significant effect on operator compliance.

Hypothesis D predicted that the false-alarm prone automation would damage the systems monitoring task by reducing operator compliance. The data agreed strongly with this hypothesis, as both the SF detection rates and response times suffered relative to the perfectly reliable automation condition, and even dropped far below Baseline performance. Although the data in Table 1 show that operators were inclined to agree with the automation when it correctly detected a system failure, the increased response times suggest that this agreement was only after double-checking the raw data. When the automation presented a false alarm, operators incorrectly agreed only a third of the time. These two factors indicate low operator compliance.

Importantly, FA-prone automation also adversely affected operator reliance, as predicted by Hypothesis E, and confirming what Wickens et al. (2005) and Dixon and Wickens (in press) suggested based on trends seen in their data. When the automation was silent, operators in the false-alarm condition should have completely ignored the systems monitoring gauge and focused their entire attention on the tracking task. Instead, the data revealed that the tracking task performance in the false-alarm condition was not only worse than the reliable automation condition, but was also worse than Baseline. This implies that the reliance-compliance constructs may not be entirely independent of each other.

Thus, our Hypothesis F, that the FA-prone condition would be more harmful to overall performance relative to the miss-prone condition, proved to be correct both qualitatively and quantitatively. First, the FA-prone automation adversely affected both operator compliance and reliance, while the miss-prone automation only appeared to reduce operator reliance. Second, FA-prone automation hurt performance more on the automated task than did miss-prone automation, (e.g., the "cry wolf" effect) and hurt performance (both speed and accuracy) at least as much as miss-prone automation on the concurrent task. The current data provide convincing evidence that automation false alarms not only produce qualitatively different effects on operator trust than do automation misses, but that they are also quantitatively more harmful to performance than misses.

There are at least three potential explanations for why automation false alarms also affect operator reliance. First, the false-alarm prone condition had a considerably larger number of discrete attention-grabbing events than any other condition, which could be particularly disruptive due to a form of auditory onset preemption (Spence & Driver, 1997). It may be possible to negate the reduction in operator reliance by equalizing the number of attention-grabbing events across conditions.

Second, automation false alarms are often more salient than automation misses (Maltz & Shinar, 2003). When an automation false alarm occurs, the operator can immediately detect the error, while an automation miss might not be noticed until the end of the trial, if at all. Because false alarms are so salient, and often annoying, it may be that they affect the operator's global trust such that the operator comes to believe that the automation is simply error prone and does not distinguish between the two types of errors (e.g. Wickens, Dixon, & Johnson, 2005). Importantly, designers need to be aware that operators may be more affected by the perceptual salience of information than by the rational consequences of the information. Therefore, designers should avoid the assumption that operators will notice information simply because it has high consequences. It may be that they instead notice such information due to high salience.

A third potential explanation may have to do with attentional preemption of the automation errors. That is, operators are previewing the system gauges prior to an alert in order to gain some understanding of the current status of the display so that when an alert sounds, they will already have some information with which to make a quicker decision. However, the current data reveals that the decision-making time is actually longer than in other conditions, so it may be that when the alert sounds, the operator spends even more time double-checking the raw data behind the automation to make sure of the decision. This confusion could potentially lead to longer response times with more incorrect decisions.

More research needs to be done to determine whether these three explanations, which are not necessarily in opposition to each other, are viable explanations for why automation false alarms appear to affect operator reliance. Further research also needs to be done to determine whether the two types of automation errors affect different cognitive processes, or whether a single-process model is sufficient in explaining the data. Our data has suggested that the two constructs are not entirely independent of each other; however, this single dissociation in the data is not sufficient to reject a single-process model (Dunn & Kirsner, 1988).

By expanding on the findings from previous studies, and providing a more sensitive analysis of the qualitative differences between automation false alarm and misses, the implications of the current study allow us to generalize beyond that of the specific UAV paradigm. Subsequently, the current data allow designers of automated systems to more accurately weigh the impact of automation false alarms and misses on operator performance when deciding where to set the bias threshold in future systems.

Acknowledgments

This research was sponsored by ARMY MAD 6021.000-01. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the Army. The authors wish to acknowledge the support of Ron Carbonari (in developing the simulation), as well as the helpful comments of three anonymous reviewers.

References

Bliss, J. (2003). An investigation of alarm related accidents and incidents in aviation. *International Journal of Aviation Psychology, 13*(3), 249-268.

Dixon, S.R. & Wickens, C.D. (in press). Automation Reliability in Unmanned Aerial Vehicle Flight Control: Evaluating a Model of Automation Dependence in High Workload. *Human Factors*.

Dunn, J.C. & Kirsner, K. (1988). Discovering functionally independent mental processes: The principle of reversed association. *Psychological Review, 95*(1), 91-101.

Green, D. M., & Swets, J. A. (1988). *Signal detection theory and psychophysics*. New York: Wiley.

Levinthal, B. & Wickens, C.D. (2005). *Supervising Two Versus Four UAVs With Imperfect Automation: A Simulation Experiment*. (AHFD-05-24/MAAD-05-7). Savoy, IL: University of Illinois, Aviation Human Factors Division.

Maltz, M., & Shinar, D. (2003). New alternative methods in analyzing human behavior in cued target acquisition. *Human Factors, 45*(2), 281-295.

Meyer, J. (2001). Effects of warning validity and proximity on responses to warnings. *Human Factors, 43*(4), 563-572.

Meyer, J. (2004). Conceptual issues in the study of dynamic hazard warnings. *Human Factors*, *46*(2), 196-204.

Spence, C., & Driver J. (1997). Audiovisual links in attention: implications for interface design. In D. Harris (Ed), *Engineering psychology and cognitive ergonomics.* Hampshire UK: Ashgate.

Wickens, C.D. & Dixon, S.R. (in press). Is There a Magic Number 7 (to the Minus 1)? The Benefits of Imperfect Diagnostic Automation: A Synthesis of the Literature. *Theoretical Issues in Ergonomics Science*.

Wickens, C.D., Dixon, S.R., Goh, J. & Hammer, B. (2005). Pilot dependence on imperfect diagnostic automation in simulated UAV flights: an attentional visual scanning analysis. *In Proceedings of the 13th Annual International Symposium of Aviation Psychology*.

Wickens, C.D., Dixon, S.R., & Johnson, N.R. (2005). *UAV Automation: Influence of Task Priorities and Automation Imperfection in a Difficult Surveillance Task.* (AHFD-05-20/MAAD-05-6). Savoy, IL: University of Illinois, Aviation Human Factors Division.